G8-Memorization-1 - Associative Memory in Iterated Overparameterized Sigmoid Autoencoders

Omar Faig Orujlu - 03750822
supervisor : Pascal Esser

August 13, 2023

1 Extension 1 - Lipschitz and Non-Lipschitz activation functions

In the paper , having non-polynomial Lipschitz function was one of the main requirements for a point x to be an attractor. That's why I tested the same network with different activation functions. I have repeated first experiment(Figure 2 in the paper) with $Tanh$ and $ReLU$ activation functions. As stated in the paper, the network should deliver similar results with sigmoidal activation functions, which is the case with Tanh. In the left graph it can be observerd that as the number of layers increase the difference between largest eigenvalues of initial and trained Jacobian gets smaller and smalller, indicating attractor behavior of the network. Biggest difference happens as expected in case of 2 layer networks.

In case of ReLU the results are different and difficult to interpret. The difference between largest eigenvalues of inital and trained jacobians is relatively small in 2 and 3 layer case however, in 4 layer network the difference is quite high suggesting attractor formation fails.

From these given two graphs the importance of Lipschitz activation functions becomes obvious for having attractor behavior in the network.

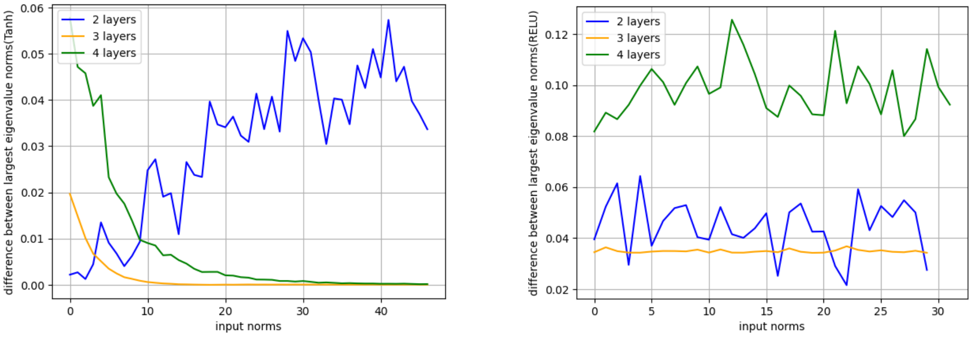


Figure 1: Lipschitz and Non-Lipschitz activation functions(Tanh on the left, ReLU on the right side)

2 Extension 2 - NTK Limit

As stated in the paper attractor formation happens only in NTK limit. For testing it and seeing the effect of NTK limit on the attractor formation the first experiment(Figure 2 in the paper) is repeated with hidden size 50 instead of 1000 with other network variables and training values being the same.

Here in 3 and 4 layer networks the difference is relatively small and at higher input norms it gets bigger. However , In 2 layer case it is 2 times higher than the original experiment.

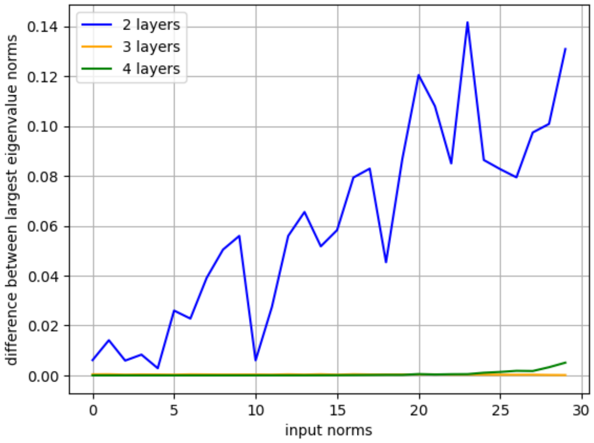


Figure 2: experiment 1 with hidden size 50

3 Extension 3 - Attractor vs Generalization

In the conclusion section of the paper the authors suggest that the relation between attractor formation and generalization of deep learning models needs to be researched. Based on this idea I conducted 5 experiments with the given parameters.

Input Norm	Hidden Size	Training Points	Validation Points
15	1000	100	30
15	1000	50	20
15	100	100	30
15	50	100	30
mixed	1000	100	30

Table 1: Experimental Setup

In the Figure 3 the validation and training loss of a network is compared. Here the right graph is zoomed-in version of the left graph. The network has hidden size of 1000 and training and validation points have norm of 15. In the right graph it can be observed that the difference between loss curves is around 0.1 till the end of training.

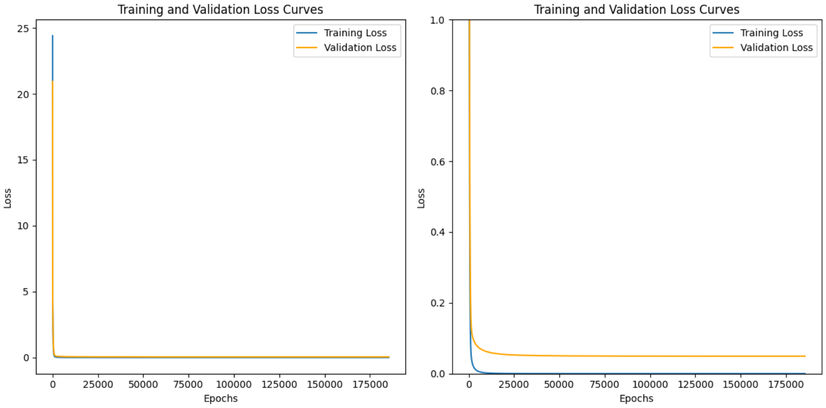


Figure 3: Hidden size 1000, norm radius 15 and 100 training points

In figure 4 the same experiment as in figure 3 is repeated with 50 training points instead of 100. Here it can be observed that the difference between the curves is almost 2 times higher than the one in Figure 3.

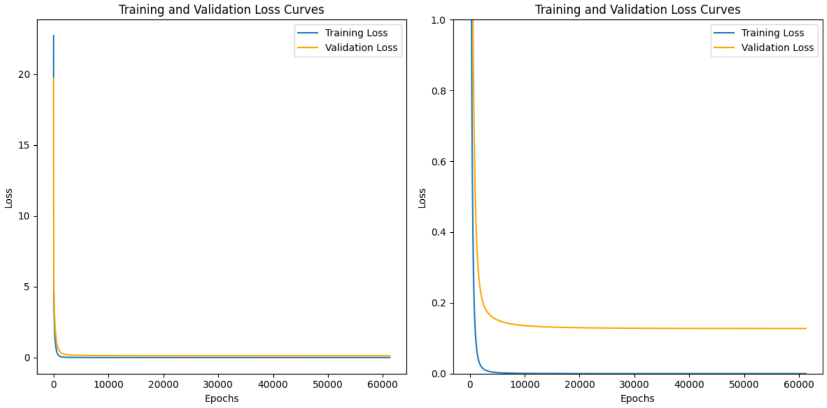


Figure 4: Hidden size 1000, norm radius 15 and 50 training points

In the figure 5 the main differnce to figure 3 is the size of hidden layers. Here the hidden layer has size of 100 instead of 1000. Here a bigger difference between 2 loss curves can be observed.

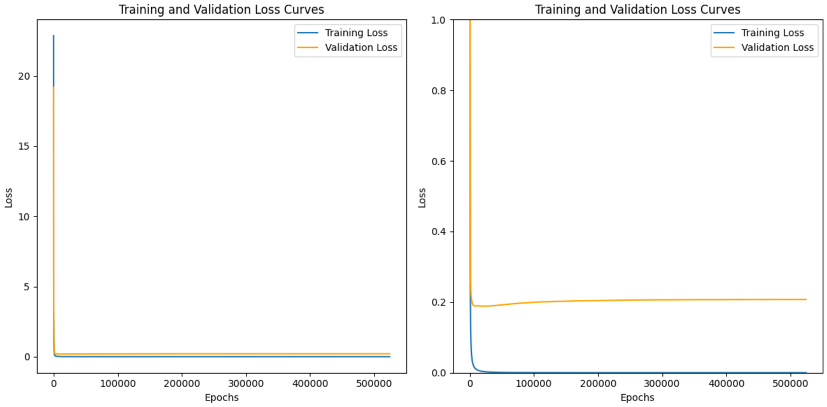


Figure 5: Hidden size 100, norm radius 15 and 100 training points

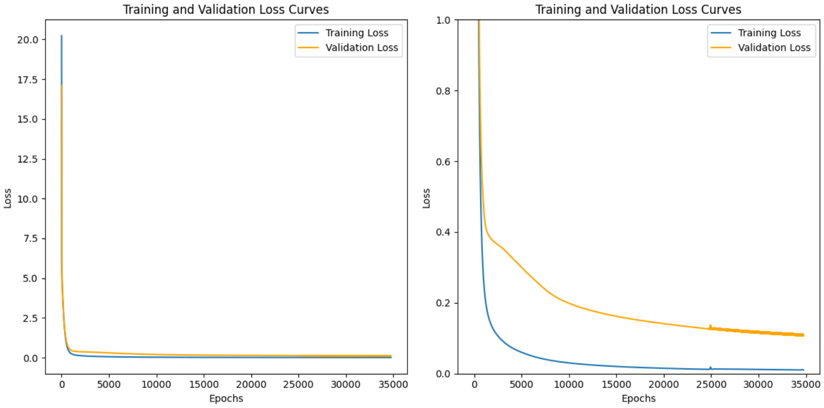
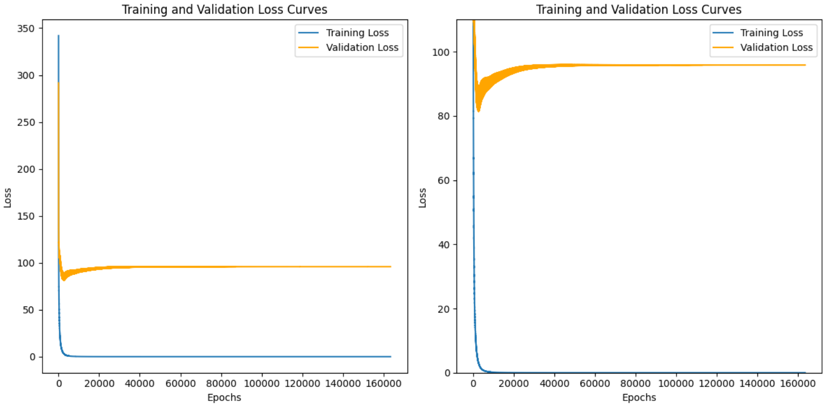


Figure 6: Hidden size 50, norm radius 15 and 100 training points and loss threshold is 10^{-2}(Non-attractor)

The figure 6 demonstrates the non-attractor behavior for comparison. Here the loss threshold for trainig is 10^{-2} instead of 10^{-7} and hidden size is 50. Here it can be observed that the difference slowly decreases and at the end plateaus.

In the last experiment of this extension mixed input norm(from 2 to 1000 randomly sampled) are trained. Despite the fact the network trains until defined loss threshold (10^{-7}) the validation does not decrease and the generalization gap is big, indicating that the network could not generalize.



4 Conclusion

For testing defined conditions and extending ideas I conducted 3 extensions with multiple experiments. Based on the experiment 1 and 2 we can support(prove) the conditions from the paper for having attractor behavior.In the 3rd extension the relationship between generalization and attractor behavior is explored. Based on 3rd extension I can say that the attractor networks do not generalize well however, additional research in this direction is necessary for proving it.